значимости. Стремление к собственной значимости - главный фактор, определяющий источник творческой активности человека, его силу и особенности. Не последнее место занимают социокультурные условия (семья, ближайшее окружение, престиж профессии и др.). Стремление к значимости собственной личности можно реализовать самым разным образом в зависимости от внешних и внутренних условий, социального уклада, жизненных обстоятельств, склонностей, способностей.

В творчестве человек предельно собран и целостен, он полностью посвящает себя служению делу, что выражается в повышенной творческой, познавательной и интеллектуальной активности. Творческий человек – это не только тот, кто стремится что-то изобрести или делает что-то новое и оригинальное, а более всего тот, кто стремится овладеть собственным поведением и собственной психической деятельностью. Представляет интерес, как процесс творчества протекает у взрослого человека, каждый ли человек считает, что он реализовал свой творческий потенциал, доступен ли этот процесс в любом возрасте, для каждой личности или необходимо своевременно мобилизовывать творческий потенциал. Всё это вопросы, требующие дальнейшего изучения.

СПИСОК ЛИТЕРАТУРЫ

1. *Ницше Ф.* По ту сторону добра и зла. Т. 2. М.: Мысль, 1990.
2. *Шопенгауэр А.* Об интересном. М.: Олимп, ООО «Издательство АСТ – ЛТД», 1997.
3. *Юнг К.Г.* Сознание и бессознательное. СПб.: Университетская книга, 1997.

В.М. Воронин, Л.И. Логвиненко

ИССЛЕДОВАНИЕ ВОСПРИЯТИЯ И ПОНИМАНИЯ ЕСТЕСТВЕННОЙ И СИНТЕЗИРОВАННОЙ РЕЧИ

Результаты проведения модифицированного райм теста (МРТ) и экспериментов по идентификации слов естественной и синтезированной речи показывают, что синтезированная речь несколько менее разборчива, чем естественная [1]. Кроме того, эксперименты показали, что, когда синтезированная речь становится все менее разборчивой, слушатели начинают все больше опираться на лингвистические правила и на ограничения круга возможных ответов, что помогает идентификации слов. Однако эксперименты не объясняют различий в восприятии естественной и синтезированной речи, их роль состояла просто в выявлении и описании этих принципиальных различий по идентификации слов.

1. Принятие лексических решений и латентность называния

Чтобы изучить различия в перцептивной обработке естественной и синтезированной речи, была выполнена серия экспериментов по измерению времени для распознавания слушателем слов и словоподобных сочетаний звуков, произносимых диктором и системой «текст – речь». Проводя эти эксперименты, мы ставили перед собой цель узнать, сколько времени нужно слушателю, чтобы идентифицировать одиночное слово, и как на процесс распознавания слов влияет качество акустико–фонетической информации в сигнале. Для того, чтобы измерить продолжительность процесса распознавания, мы, как и в работах D.B. Pisoni и др. [3,6], воспользовались задачей принятия лексических решений. В экспериментах использовался русско–язычный синтезатор Speakiug Mouse Home и англо–язычный синтезатор Speech Synthezier 5.0. Слушателю в каждой экспериментальной попытке предъявлялось либо одиночное слово, либо стимульная словоподобная единица. Каждый слушатель должен был как можно быстрее и точнее классифицировать стимульную единицу как «слово» и «неслово», нажимая одну из двух кнопок на блоке ответов, соединенном с персональной ЭВМ. Примеры русскоязычных стимулов приведены в табл. 1, англоязычных - в табл. 2.

Слушатели реагировали на *произнесенные диктором* слова русского языка (978 мс) и «неслова» (1008 мс) немного быстрее, чем на слова и «неслова», *синтезированные машиной* (1096 и 1203 мс соответственно). В среднем время реакции на синтезированную речь на 137 мс превышало время реакции на естественную речь. Полученные данные указывают на существование двух важных различий в восприятии естественной и синтезированной речи. Во-первых, восприятие синтезированной речи требует больше умственных усилий, чем восприятие естественной речи. Во-вторых, различия в латентности наблюдались в равной степени как для слов, так и для «неслов», и поэтому они не зависят от лексического статуса стимула. Отсюда следует, что дополнительные усилия при обработке информации, по–видимому, связаны с процессом выделения из сигнала акустико–фонетической информации, а не с процессом идентификации слов в лексиконе. Резюмируя, можно сказать, что совокупность полученных результатов позволяет предположить, что перцептивные процессы по расшифровке синтетической речи требуют больших умственных «усилий» или ресурсов, чем процессы расшифровки естественной речи.

Как и в эксперименте с принятием лексических решений, слушатели заметно больше времени затрачивали на называние синтезированных тестовых стимулов, чем на называние естественных стимулов. Валидное различие опять–таки наблюдалось как для слов, так

и для «неслов». Результаты экспериментов с пониманием показывают, что дополнительное время, требуемое для обработки синтезированной речи, зависит от типа ответа слушателя, поскольку близкие результаты получаются как при ответе путем нажатия кнопки, так и при устном ответе.

Таблица 1

Примеры русско–язычных стимулов для принятия лексических решений

«слово»	«неслово»
агроном	атор
бор	борж
ванна	вериг
гиря	гирс
код	крин
лицо	линат
томаты	тесло
сирень	сагра
окно	онол

Таблица 2

Примеры англо–язычных стимулов для принятия лексических решений

«СЛОВО»	«НЕСЛОВО»
PROMINENT	PRADAMENT
BAKED	BEPT
TINY	TADGY
GLASS	GEEP
PARENTS TOLD	PEEMERS TAVED
BLACK	BAEP
CONCERT	CAELIMPS
DARK	DUT
BABBLE	BURTLE
CRITIC	CRAENICK
BOUGHT	BUPPED
PAIN	POON
GORGEOUS	GAETLESS
COLORED	COOBERED

Полученные результаты показывают, что начальные этапы расшифровки синтезированной речи занимают больше времени, чем при расшифровке естественной речи. Для того, чтобы выяснить характер и степень различия процесса расшифровки естественной и синтезированной речи, было выполнено несколько дополнительных исследований.

2. Смешение слогов из согласных и гласных

В объяснение большей трудности расшифровки синтезированной речи можно предложить несколько гипотез. Одна из них предполагает, что синтезированная речь просто аналогична «зашумленной» естественной речи. Иными словами акустико-фонетическую структуру синтезированной речи труднее декодировать, чем естественную по тем же самым причинам, по каким трудно воспринимать естественную речь, предъявляемую на фоне шума, – наличие маскирующего шума приводит к искажениям акустических признаков фонем или ухудшению акустического качества фонем. В соответствии с этой точкой зрения синтезированная речь по параметрам близка к естественной речи, однако по сравнению с естественной является менее четко выраженной. Согласно другой гипотезе, синтезированная речь вовсе не напоминает «зашумленную» или нечеткую естественную речь, а может рассматриваться как «перцептивно новая», обедненная по сравнению с естественной речью. С этой точки зрения, синтезированная речь принципиально отличается от естественной как в количественном, так и в качественном отношении, поскольку слабо представлены (или вообще отсутствуют) важные акустические признаки, по которым производится распознавание. Разговорный язык является богатым в структурном отношении и избыточен на всех лингвистических уровнях. В частности, естественная речь весьма избыточна на уровне акустико–фонетической структуры.

Каждое фонетическое отличие в естественной речи слагается из множества акустических признаков, которые меняются в зависимости от контекста, темпа речи и диктора. При синтезе речи по правилам обычно используется только часть этих признаков, входящих в фонетические правила реализации. В результате, как считают сторонники второй гипотезы, некоторые фонетические отличия могут иметь лишь минимальное число признаков, что если в разных фонетических понятиях не все признаки одинаково важны, то одного единственного признака при восприятии может оказаться недостаточно, чтобы передать некоторое конкретное фонетическое отличие во всех встречающихся высказываниях. Более того, опора на минимальный набор признаков при синтезе речи, по их мнению,

может оказаться роковой для восприятия такой речи, если некоторое фонетическое отличие будет синтезировано неправильно или если оно будет маскироваться окружающим шумом.

Исходя из двух указанных выше гипотез относительно структурных отношений между синтезированной и естественной речью, возможны два предсказания относительно характера ошибок и распределения ошибок восприятия типа смешения (то есть когда один сегмент ошибочно принимается за другой), которые будут наблюдаться при сравнении синтезированной речи с естественной. В соответствии с гипотезой о «зашумленной речи», синтезированная речь аналогична естественной речи, качество которой ухудшено из-за добавления шума. Поэтому характер ошибок восприятия типа смешения, которые будут наблюдаться при восприятии синтезированной речи, должен быть весьма сходным с характером ошибок при прослушивании естественной речи на шумовом фоне. В отличие от этого гипотеза «обедненной речи» утверждает, что акустико-фонетическая структура синтезированной речи не столь богата сегментными признаками и не столь избыточна, как естественная речь. В соответствии с ней при восприятии синтезированной речи должны наблюдаться два типа ошибок смешения. Если акустические признаки, используемые при синтезе фонетического сегмента, не являются достаточно характерными, то должно происходить смешение тех сегментов, имеющих минимальное число признаков, которые фонетически схожи. Этот тип ошибок аналогичен ошибкам, которые предсказывает гипотеза зашумленной речи, поскольку путаница при восприятии зашумленной естественной речи также зависит от акустико-фонетического сходства сегментов [4]. Однако справедливость одной из двух гипотез можно установить по наличию второго типа ошибок, которые предсказывает только гипотеза обедненной речи. Если минимальный набор акустических признаков, используемый при синтезе фонетически отличающихся сегментов, некорректен или же вводит слушателя в заблуждение в результате плохо сформулированных фонетических правил реализации, то будут наблюдаться также ошибки, которые не связаны с номинальным акустико-фонетическим сходством тех сегментов, которые слушатель путает при восприятии. Напротив, эти ошибки должны полностью определяться перцептивной интерпретацией слушателем тех признаков, которые вводят в заблуждение. Таким образом, фонетический характер ошибок, наблюдаемых при восприятии синтезированной речи, должен быть совершенно иной, нежели тот, которого можно было бы ожидать

исходя из акустико-фонетического сходства с естественной речью. Для того, чтобы проверить предсказания, вытекающие из двух указанных гипотез, мы, как и авторы работы [5], провели анализ ошибок восприятия, возникающих при использовании в качестве стимулов набора из 48 слогов «согласная – гласная (СГ)», производимых голосом или синтезатором. Эти слоги были получены путем комбинации гласных / i, a,u / и согласных / b, d, g, p, t, k, m, n, z, l, w, j, s, f, z, v /. В наших экспериментах использовался, как уже отмечалось выше, англоязычный синтезатор Speech Synthezier 5.0.

Естественные слоги из гласных и согласных произносил диктор-мужчина. В экспериментах, описанных в работе [5], синтезированные слоги генерировались тремя системами «текст – речь»: Votrax Type-`N`-Talk, Speech Plus Prose-2000 v.2.1 и DECtalk v.1.8 фирмы Digital Equipment. Чтобы оценить характер ошибок восприятия при естественной речи, соответствующие слоги предъявлялись слушателям для прослушивания при четырех отношениях сигнал / шум (С/Ш): +28, 0, – 5, и –10 дБ.

Результаты после усреднения по слогам, содержащим все три гласные и полученным с помощью трех синтезаторов, в работе [5] показали, что наиболее разборчива естественная речь с отношением С/Ш +28 дБ (96.6% правильных ответов). Следом идет синтезированная речь, получаемая на системе DECtalk (92.2% правильных ответов), затем синтезированная речь системы Prose-2000 (62.8% правильных ответов). Самые плохие результаты получены для синтезатора Type-'N'-Talk (27% правильных ответов). Особый интерес представляли результаты более детального анализа ошибок, который показал, что характер ошибок восприятия типа смешения зачастую совершенно разный для естественной и синтезированной речи. Например, для синтезатора DECtalk 100% ошибок при идентификации сегмента /r/ оказались ошибками смешения с /b/, хотя такой тип ошибки никогда не встречался при восприятии естественной речи с отношением С/Ш +28 дБ. Даже при самом «плохом» отношении С/Ш (– 10 дБ), когда разборчивость естественной речи в шуме была намного хуже, чем синтезированной речи системы DECtalk в условиях без шума (29.1% правильных ответов против 92.2%), этот тип ошибки составил всего 3% от общего числа ошибок для данного сегмента.

Чтобы точнее оценить характер ошибок типа смешения при восприятии сегментов, в работе [5] авторы сопоставляли две «матрицы смешения» – для конкретной системы «текст – речь» и для естественной речи, предъявляемой с отношением С/Ш, которое обеспечивало

примерно такую же правильность идентификации, что и в случае синтезированной речи. Они провели сравнение матриц смешения для речи, получаемой на синтезаторе Prose-2000, и для естественной речи, имеющей отношение С/Ш 0 дБ, а также матриц смешения для речи, получаемой на синтезаторе Vortax, и для естественной речи, предъявляемой при отношении С/Ш –10 дБ. Изучение вкладов ошибок при восприятии разных типов сегментов (взрывных, носовых, плавных / полугласных, фрикативных и пр.) в общее число ошибок показало, что в случае естественной речи большинство ошибок при идентификации взрывных связано с называнием слушателями других взрывных согласных. В отличие от этого при восприятии речи, синтезированной на системе Prose-2000, распределение ошибок оказалось более равномерным – при ошибочных ответах назывались как взрывные, так и плавные / полугласные и фрикативные. Иными словами, по сравнению с естественной речью, предъявляемой в шуме, при восприятии синтезированной речи, получаемой на системе Prose-2000, в большей мере обнаруживается наличие разных типов ошибок. Таким образом, разный характер ошибок, наблюдаемых при восприятии речи, создаваемой синтезатором Prose-2000, и естественной речи, позволил предположить авторам работы [5], что ошибки в случае синтезатора Prose-2000 могут объясняться «фонетическими лжепризнаками», а не истинным фонетическим смешением.

Сравнение естественной речи, предъявляемой с отношением С/Ш — 10 дБ, и речи, синтезированной системой Votrax, показало, что в этом случае характер ошибок при идентификации взрывных более схож. В самом деле, сравнение ошибок идентификации для естественной речи при отношении С/Ш 0 дБ и –10 дБ дает совершенно такой же результат, как сравнение речи, синтезированной на Votrax с естественной речью. Таким образом, по крайней мере, при восприятии взрывных согласных ошибки смешения для речи, синтезированной на системе Votrax, по-видимому, объясняются акустико-фонетическим сходством смешиваемых сегментов, как и в случае зашумленной речи. Однако следует подчеркнуть, что в целом правильность распознавания синтезированной речи системы Votrax была весьма низкой. Поэтому ошибки могут быть сходными по той причине, что правильность распознавания приближается к уровню случайного угадывания.

Совсем иной характер ошибок у авторов работы [5] наблюдался при восприятии плавных и полугласных. Как оказалось, распределение ошибок при восприятии плавных и полугласных было сходным для естественной речи и для речи, синтезированной на системе Prose-2000. Однако для тех же самых фонем, синтезированных на системе Votrax,

распределение ошибок было совершенно иным, чем для естественной речи. Наибольшее число ошибок при восприятии речи, синтезированной на системе Votrax, происходило из-за смешения плавных и полугласных с взрывными согласными, тогда как при восприятии естественной речи случаев смешения с взрывными отмечалось немного. Таким образом, при восприятии плавных и полугласных, синтезированных на системе Prose-2000, ошибки объясняются в основном акустико-фонетическим сходством, тогда как ошибки при восприятии речи, синтезированной на Votrax, судя по всему, объясняются фонетическими «лжепризнаками».

Подводя итоги анализа ошибок, авторы [5] делают выводы о том, что предсказания, вытекающие из гипотезы о зашумленной речи, не подтвердились. При восприятии синтезированной речи наблюдалось два типа ошибок. Часть ошибок при идентификации согласных объясняется акустико-фонетическим сходством смешиваемых сегментов. Характер остальных ошибок можно объяснить только фонетическими лжепризнаками – это такие ошибки, когда акустические признаки, используемые при синтезе речи, в конкретном контексте дают ложный сегмент.

Результаты экспериментов по смешению слогов из согласной и гласной, считают Pisoni и его соавторы, свидетельствуют в пользу того, что различия в восприятии естественной и синтезированной речи объясняются главным образом различными акустико–фонетическими свойствами сигналов. Они также считают, что дальнейшее подтверждение этого положения получено при исследовании восприятия естественных и синтезированных слов по методу стробирования [3].

В эксперименте слушателям для идентификации предъявляли короткие сегменты слов либо естественной, либо синтезированной речи. В первой экспериментальной попытке с предъявлением конкретного слова для идентификации предлагался отрезок сигнала длительностью 50 мс (с начала слова). В последующих попытках длительность сигнала увеличивалась шагами по 50 мс, так что во второй попытке для идентификации предъявлялся отрезок слова длительностью 100 мс, в третьей – 150 мс и т.д., пока не предъявлялось полностью все слово. Авторы работы [3] выяснили, что в среднем слова естественной речи могут быть идентифицированы, когда на слух воспринимается 67% всего слова. Для правильной идентификации синтезированных слов слушателю необходимо было слышать 75% всего слова. Результаты экспериментов со «стробированием», по мнению Pisoni, четче показывают, что акустико-фонетическая структура синтезированных слов несет меньше информации (в единицу времени), чем акустико-фонетическая структура естественной речи.

Наши эксперименты с англо–язычным синтезатором Speech Synthezier 5.0. показали, однако, что распределение ошибок при восприятии синтезированной речи, как и в случае естественной речи, не было равномерным, то есть ошибки при идентификации, например, взрывных связаны с называнием слушателями других взрывных согласных. Аналогично характер ошибок при восприятии плавных и полугласных было сходным как для естественной речи, так и для синтезированной речи. Это справедливо при любом соотношении С/Ш, то есть 28 дБ, 0 дБ, –5дБ.

Мы не проводили экспериментов со стробированием сигнала, но можно прогнозировать, что результаты будут отличаться от результатов, полученных в работе [3], иными словами можно предположить, что средняя длительность слова, необходимая для правильной идентификации, вряд ли будет различна как для естественной, так и для синтетической речи.

3. Выводы: Перцептивная расшифровка

Рассматриваемые в совокупности результаты наших исследований явно свидетельствуют о том, что расшифровка акустико-фонетической структуры синтезированной речи более сложна, чем расшифровка естественной речи, и требует больших умственных усилий и объема памяти. Одним из оснований для такого вывода являются данные о том, что распознавание слов и «неслов» синтезированной речи требует большего времени обработки, чем естественной речи. Это говорит о том, что главная трудность при распознавании связана с выделением фонетической информации, а не с самим по себе распознаванием «неслов», поскольку один и тот же результат получен как для слов, так и для «неслов». Главный вывод на основе полученных при изучении смешения слогов из гласных и согласных заключается в том, что в отличие от экспериментов американских авторов не выявлено заметной разницы в восприятии акустико-фонетической структуры синтезированной и естественной речи. В сравнении с естественной речью синтезированную речь можно рассматривать не как фонетически обедненную, а как аналог зашумленной естественной речи.

В целом, вся совокупность полученных результатов позволяет предположить, что различия в продолжительности обработки естественной и синтезированной речи, скорее всего, появляются на этапах обработки, связанных с выделением из речевого сигнала основной акустико-фонетической информации, то есть на начальных этапах самого процесса распознавания, а не на более поздних, познавательных уровнях,

связанных с поиском или с извлечением слов, хранящихся во внутреннем (связанном с мозговыми структурами) словаре.

Подтверждение нашей гипотезы о том, что синтезированную речь можно считать аналогом зашумленной естественной речи объясняется тем, что в отличие от экспериментов американских авторов мы использовали более совершенный англо-язычный синтезатор.

Это положение имеет, на наш взгляд, существенное методологическое значение для когнитивной психологии и в целом для когнитивной науки. Оно заключается в том, что чисто человеческая функция – функция речеобразования выполнена технической системой таким образом, что восприятие синтезированного речевого сигнала по пространственным (а значит и спектральным) и временным характеристикам адекватно восприятию естественного речевого сигнала.

С другой стороны, принятие упомянутой выше гипотезы предоставляет простой и эффективный способ оценки качества того или иного синтезатора речи. Действительно, эффективность синтезатора в этом случае определяется соотношением сигнала к шуму (С/Ш) для предъявляемой естественной речи, которое обеспечивает примерно такую же правильность идентификации, что и генерируемая синтетическая речь.

$$(1) \quad Э = С/Ш$$

Отсюда, если сравнивать, например, два синтезатора, то соотношение их эффективностей будет выражаться формулой:

$$(2) \quad Э_1/Э_2 = (С/Ш)_1/(С/Ш)_2$$

Результаты, полученные в экспериментах с принятием лексических решений и с называнием, также показывают, что даже при сравнительно высокой правильности распознавания в задаче распознавания отдельных слов синтезированная речь требует несколько большего времени обработки на когнитивных уровнях, чем естественная. В этих опытах, однако, слушатели решали относительно простые и прямые задачи. Как мы отметили в начале, конкретные требования и условия задачи в экспериментах по восприятию почти всегда оказывают влияние на скорость и точность ответов слушателя. Следующая серия экспериментов, к описанию которой мы перейдем, поставлена так, что помимо требований, предъявляемых при выявлении акустико-фонетических свойств синтезированной речи, к слушателям предъявляются дополнительные требования познавательного характера.

4. Требования к объему памяти при восприятии

Исходя из работ по избирательному вниманию человека, можно предположить, что ограничивающим фактором для когнитивных процессов является объем кратковременной, или

оперативной, памяти [7]. Так, любой процесс восприятия, использующий кратковременную память, может мешать процессу принятия решений с перцептивной обработкой и с другими операциями следующих, когнитивных уровней анализа. Если при восприятии синтезированной речи требуется больший объем кратковременной памяти, чем при восприятии естественной речи, то в случае практических применений синтезированной речи, когда при распознавании речевого сообщения необходимо выполнять и другие сложные когнитивные операции, могут возникнуть определенные трудности. С этой целью была выполнена серия экспериментов по определению загрузки кратковременной памяти при обработке синтезированной речи. В одном из таких экспериментов испытуемым в каждой попытке предъявляли для запоминания два различных списка. Первый список состоял из цифр, *визуально* предъявляемых с экрана монитора. В некоторых попытках цифры отсутствовали, в других предъявлялись либо три, либо шесть цифр. После показа списка испытуемым на слух предъявляли десять слов естественной или синтезированной речи. После предъявления на слух испытуемые, согласно инструкции, должны были записать все визуально предъявленные цифры в порядке их показа, а также все слова, которые им удалось запомнить на слух. Во всех трех условиях зрительного предъявления (отсутствие цифр, три и шесть цифр) запоминание слов естественной речи было лучше, чем синтезированных слов. Кроме того, по мере того, как удлинялся список цифр, запоминание слов как естественной, так и синтезированной речи, ухудшалось. Другими словами, увеличение ряда цифр, хранимых в кратковременной памяти, ухудшает запоминание слов, предъявляемых на слух. Однако наиболее важным результатом оказалось обнаружение взаимосвязи между характером предъявляемой речи (синтезированая или естественная) и числом цифр (три или шесть), Эта взаимосвязь была выявлена по количеству испытуемых, которые могли в правильном порядке назвать все предъявленные цифры. По мере удлинения списка цифр меньшее количество испытуемых могло назвать цифры в случае синтезированной речи (по сравнению с естественной речью). Таким образом, при увеличении списка цифр из-за стремления запомнить визуально предъявленные цифры восприятие синтезированной речи нарушается больше, чем восприятие естественной речи. Этот результат свидетельствует, что для обработки синтезированной речи требуется больший объем кратковременной памяти, чем для обработки естественной речи. В итоге, поскольку восприятие синтезированной речи требует большего объема памяти, чем восприятие естественной речи,

следует ожидать, что синтезированная речь будет сильнее мешать другим когнитивным процессам.

Чтобы проверить этот вывод, мы провели еще один эксперимент, в котором испытуемые должны были запоминать списки из десяти слов. Каждый из списков был составлен либо полностью из синтезированных слов, либо полностью из слов естественной речи. Испытуемый должен был воспроизвести слова из списка в том порядке, в каком они предъявлялись. Как и в предыдущем эксперименте, слова естественной речи запоминались лучше, чем синтезированные слова. Однако более тщательный анализ выявил связь запоминания с положением слов в списке. Первые из воспринятых на слух синтезированных слов запоминались гораздо менее точно, чем слова естественной речи в начале списков. Это говорит о том, что при запоминании на слух синтезированных слов, те слова, которые произносятся позже, «мешают» запоминанию тех, которые были произнесены раньше и которые активно повторяются испытуемым при запоминании. Именно такого результата и следовало ожидать, если процесс расшифровки синтезированных слов требует большего объема кратковременной памяти.

Результаты по запоминанию и воспроизведению в порядке предъявления списков слов естественной и синтезированной речи подтверждают вывод, следующий из работы по принятию лексических решений: восприятие синтезированной речи требует больших ресурсов при обработке, чем восприятие естественной речи. Расшифровка при восприятии синтетической речи требует большей емкости памяти и может, в свою очередь, влиять на другие когнитивные процессы, требующие активных ресурсов внимания. Предыдущие работы по оценке ограничений, вносимых памятью, при восприятии речи показали, что перевод внимания на одно речевое сообщение значительно снижает способность слушателя обнаруживать конкретные слова в других речевых сообщениях. Кроме того, несколько экспериментальных работ выявили, что перевод внимания на одно сообщение приводит к существенному снижению точности распознавания фонем в другом речевом потоке. Если рассматривать совместно результаты всех этих работ, то они указывают, что распознавание речи даже на уровне декодирования фонем требует активного внимания и определенной емкости памяти. Вследствие этого повышенные требования, которые предъявляются к системе обработки при расшифровке синтезированной речи, могут привести к существенным ограничениям на восприятие и распознавание при применении систем с речевым выходом в условиях высокой информационной загрузки или в сложной обстановке. Это

особенно справедливо в тех случаях, когда слушатель должен одновременно уделять внимание нескольким разным источникам информации.

5. Влияние тренировки и опыта работы с синтезированной речью

Мозг человека является весьма гибким устройством для обработки информации. После специальной тренировки, приобретения опыта и достаточной практики слушатели способны преодолеть некоторые из ограничений, которые мы наблюдали в наших предыдущих экспериментах. В самом деле, ряд исследователей [2] сообщали о быстром улучшении распознавания синтезированной речи в ходе своих экспериментов. Это улучшение является, по-видимому, результатом того, что испытуемые научились более эффективно обрабатывать акустико-фонетическую структуру синтезированной речи. Однако возможно и другое объяснение – что наблюдаемое повышение степени понимания синтезированной речи связано просто с лучшим овладением методикой эксперимента, а не с истинным улучшением перцептивной обработки синтезированной речи. Для того, чтобы определить роль каждой из этих двух причин, мы провели эксперимент, в котором влияние тренировки было отделено от улучшения распознавания синтезированной речи.

Трем группам слушателей (студенты 5–го курса университета) в ходе эксперимента по восприятию синтезированной речи давали начальный тест (протест) в первый день и заключительный тест (посттест) на 10-й день эксперимента. Протест позволял определить исходную степень правильности распознавания синтезированной речи системы «текст – речь», Посттест, предъявляемый на 10-й день, служил для определения того, произошло ли какое-либо улучшение распознавания синтезированной речи после тренировки. Указанные три группы слушателей в период со 2-го по 9-й день эксперимента получали разные задания. Одна группа тренировалась на синтезированной речи упрощенного синтезатора [см. 1], вторая - на естественной речи, причем использовались те же самые слова, предложения и отрывки текста, что и для первой группы. Эта вторая группа была контрольной, по ней оценивалось влияние навыков овладения конкретной экспериментальной методикой. Наконец, третья группа слушателей в период со 2-го по 9-й день вообще не тренировалась. Результаты показали, что резкое улучшение распознавания произошло только в одной группе, где слушатели тренировались на синтезированной речи упрощенного синтезатора. В конце периода тренировки слушатели этой группы показали заметно более высокую степень распознавания, чем две другие группы. Например, правильность идентификации отдельных фонематически

сбалансированных слов (ФСС) улучшилась в первой группе примерно с 35% (по результатам протеста) почти до 85% (посттест). Примерно такое же улучшение наблюдалось и для всех остальных слов, применявшихся в задачах идентификации.

Эти результаты по влиянию тренировки позволяют сформулировать несколько важных выводов. Во-первых, эффекты тренировки, по-видимому, связаны с улучшением или модификацией процесса расшифровки, использующегося при распознавании слов. Ясно, что улучшение распознавания не связано с тем, что слушатели просто научились лучше выполнять различные задачи, поскольку у тех лиц, которые тренировались на естественной речи, улучшения распознавания не наблюдалось или оно было незначительным. Более того, тренировка одинаково влияла как на распознавание отдельных слов, так и на распознавание слов в составе предложений как при ограниченном, так и при неограниченном выборе. Все это позволяет с достаточным основанием предположить, что слушатели из группы, тренировавшейся на синтезированной речи, не пользовались ни запоминанием отдельных тестовых стимулов, ни выработкой специальных стратегий. Иными словами, для повышения степени распознавания в процессе тренировки они не научились использовать лингвистические значения или ограничения на задачу. Скорее испытуемые усвоили нечто такое о структурных характеристиках данной конкретной системы синтеза «текст –речь», что позволило им лучше работать независимо от решаемой задачи. Дополнительное подтверждение этого вывода дают результаты нашего эксперимента по оценке влияния тренировки. Повышение правильности распознавания было получено на новом материале, несмотря на то что слушателям в течение всего эксперимента одни и те же слова или предложения никогда не предъявлялось более одного раза. Для того, чтобы в таких условиях повысить правильность распознавания, слушатели должны были бы досконально знать всю систему правил, используемую при синтезе речи данной системой. Если бы слушатели просто стали запоминать отдельные слова или целые предложения, то не смогли бы повысить правильность распознавания в посттесте, поскольку в этом тесте также использовались новые материалы.

В дополнение к этим результатам мы также выяснили, что тренировочный эффект сохраняется даже спустя шесть месяцев после эксперимента, в течение которых слушатели не имели случая слышать синтезированную речь данного типа. Таким образом, не вызывает сомнений, что тренировка привела к достаточно стойким и долговременным изменениям процессов восприятия и расшифровки, которыми пользовались слушатели. Кроме того, весьма вероятно, что более интенсивная тренировка

привела бы к еще более стойким тренировочным эффектам. Если бы слушатели продолжали тренироваться до тех пор, пока не был бы достигнут асимптотический уровень правильности распознавания (то есть когда новые этапы тренировки приводили бы лишь к незначительному повышению правильности распознавания), то долговременные эффекты тренировки могли бы оказаться еще более стойкими.

Результаты данного эксперимента свидетельствуют, что при расшифровке синтезированной речи слушатель может модифицировать свои стратегии восприятия и что существенного повышения правильности распознавания можно достигнуть в относительно короткие сроки даже в случае синтезированной речи низкого качества.

6. Дальнейшие направления исследований

6.1 Изучение понимания речи

Большинство работ по синтезу речи системами «текст – речь» в 80-е и в начале 90-х годов в США было сконцентрировано главным образом на акустико-фонетических возможностях систем. Основное внимание уделялось улучшению сегментной разборчивости синтезированной речи. Имеющиеся в настоящее время и полученные нами данные позволяют полагать, что сейчас системы синтеза речи имеют достаточно хорошую сегментную разборчивость, приближающаяся к разборчивости речи, и есть резервы для дальнейшего увеличения. С другой стороны, мало внимания уделяется оценке понимания речи (в более общем смысле) на слух. Во всех предыдущих работах использовались относительно грубые и низкочувствительные критерии понимания. Несмотря на это, нам удалось обнаружить небольшое, но устойчивое различие в степени понятности естественной и синтезированной речи.

Для того, чтобы полностью определить роль, которую играет опыт и обучение в восприятии и понимании речи, необходимы дальнейшие исследования. Как мы уже отмечали выше, у лиц, приобретших опыт в прослушивании синтезированной речи, возрастает степень понимания ее на слух. Необходимо выполнить дополнительные эксперименты с целью выяснения характера эффектов обучения и тренировки, а также определения того, как меняются критерии стратегии восприятия слушателей. Здесь возникает множество вопросов: какова должна быть продолжительность тренировки слушателя; можно ли достичь такой же правильности распознавания синтезированной речи, как и естественной речи; уменьшаются ли по мере тренировки требуемые ресурсы при восприятии синтезированной речи. Для ответа на эти вопросы необходимы новые, тщательно поставленные лабораторные

эксперименты с использованием более сложных и эффективных критериев восприятия и понимания речи.

6.2 Текущие оценки механизмов лингвистической обработки

Для того, чтобы понять, каковы требования в каждый момент времени при прослушивании беглой синтезированной речи, нам понадобятся те же текущие оценки, что и для реальновременных вычислительных процессов, протекающих, когда слушатели воспринимают и расшифровывают содержание беглой речи. Для получения информации о тех скрытых процессах, которыми пользуются слушатели, чтобы понять синтезированную речь, необходимо использовать задачи с фонемным и словарным управлением, в которых слушатели должны давать ответы в ходе прослушивания речи. Полезными могут оказаться также и другие психолингвистические задачи, например, задача обнаружения неправильного произношения. В этих задачах для количественной оценки когнитивных механизмов обработки информации используются данные о времени реакции слушателей.

6.3 Привыкание к синтезированной речи и концентрация внимания

При длительном прослушивании синтезированной речи слушатель зачастую испытывает трудности в сосредоточении внимания на лингвистическом содержании текста. Несмотря на то, что в наших экспериментах по оценке понимания речи выявилась достаточно высокая степень понимания смысла отрывков слушателями, у нас нет никаких доказательств, что слушатели полностью концентрировали все свое внимание на этих отрывках. Из субъективных отчетов по другим экспериментам мы можем предположить, что в процессе прослушивания продолжительных отрывков синтезированной речи слушатели то «настраиваются», то «отключаются». Быстрее ли устают слушатели при восприятии синтезированной речи, чем при восприятии естественной речи? Может ли слушатель полностью понять содержание отрывка, если он расслышал только некоторую его часть? Как распределяется внимание слушателя при прослушивании синтезированной речи по сравнению, например, с прослушиванием естественной речи, и как изменение этого распределения связано с потребностями в емкости кратковременной памяти? Все эти важные вопросы ожидают дальнейших исследований.

6.4 Субъективное оценивание и предпочтение слушателя

Помимо вопроса о качестве синтезированных речевых сигналов существуют и другие аспекты оценки синтезированной речи, связанные со склонностями и вкусами пользователей. Если человеку, который работает с той или иной системой синтеза речи по тексту, не нравится звучание этой речи или если он не доверяет тем сообщениям, которые воспроизводятся голосовым устройством, то польза, приносимая

подобным устройством, оказывается значительно сниженной. Поэтому необходим вопросник, позволяющий субъективно оценивать синтезированную речь. Некоторые предварительные данные были собраны с применением различных типов стимулов и разных синтезаторов. Выяснилось, что субъективные оценки слушателями качества речи синтезаторов, как правило, достаточно хорошо коррелируют с объективными оценками характеристик синтезаторов. Кроме того, в своих экспериментах мы установили, что степень доверия слушателей к информации, выдаваемой в синтезированной речи, положительно коррелирует с объективными оценками характеристик синтезаторов. Если пользователь никогда не сталкивался с синтезированной речью, то низкое качество синтезированной речи вызывает такой же низкий уровень доверия к сообщениям, тогда как при высоком качестве степень доверия к сообщениям больше.

6.5 Исследования в области применения аппаратуры с речевым выходом

Сейчас нет ответов на многие вопросы, связанные с практическим применением систем с речевым выходом. Нужны дополнительные исследования по текущему и будущему применению синтезированной речи. За исключением нескольких работ, где описано применение синтезированной речи в военной и деловой сферах, а также в промышленности, в большинстве работ по применению синтезированной речи просто описываются все новые и новые примеры и не предпринимаются попытки оценить их успешность, пользу или проанализировать примеры неудачного применения.

Итоги и выводы

Оценка эффективности использования систем синтеза с речевым ответом не ограничивается проведением стандартных тестов на разборчивость. Различные практические задачи будут предъявлять к операторам разные требования и налагать разные ограничения. Поэтому необходимо учитывать те пять факторов, о которых мы говорили в другой нашей работе [1], и определить, как они будут сочетаться и взаимодействовать, влияя на правильное понимание человеком синтезированной речи. В частности, следует учитывать, что перцептивные и когнитивные механизмы ограничиваются, в первую очередь, объемом кратковременной памяти. Поскольку при восприятии синтезированной речи кратковременная память сильно загружена, можно считать, что в задачах, требующих особо сильной загрузки кратковременной памяти, обработка синтезированной речи может отрицательно сказываться на

выполнении других, одновременно протекающих когнитивных операций. Разумеется, не исключена и противоположная ситуация, то есть решение задачи, сильно загружающей кратковременную память, может мешать обработке синтезированной речи. Человек не похож на ЭВМ, работающую в режиме с прерываниями, которая может сразу же дать ответ при предъявлении входного сигнала. Во время сложной когнитивной обработки информации человек может оказаться не в состоянии правильным образом отреагировать на речевой сигнал. Более того, в сложной обстановке и в критической ситуации или при решении очень сложной задачи человек может вообще не обратить внимания на сообщение, передаваемое синтезированной речью.

Если человек решает очень трудную задачу, передаваемые ему сообщения должны характеризоваться максимальной избыточностью и перцептивной различимостью. Именно в таких ситуациях структура и содержание набора сообщений становятся критическими. Когда набор сообщений упрощается (например, в случае отдельных слов-команд), следует соответственно увеличить перцептивную различимость сообщений. При восприятии отдельных слов слушатель не может полагаться на лингвистические ограничения, налагаемые синтаксисом и семантикой. Кроме того, наиболее важна различимость сообщений тогда, когда фонемный синтез имеет низкое качество, поскольку при этом минимальна избыточность акустической структуры сигнала. В результате слушатель при расшифровке речи затрачивает больше умственных усилий. Это означает, что синтезаторы речи с упрощенным синтезом должны использоваться только в тех случаях, когда решаются не очень трудные задачи.

Кроме того, следует отдавать себе отчет в том, что, исходя из результатов традиционного теста МРТ с ограниченным выбором, нельзя прогнозировать, какова будет реакция на синтезированную речь при решении практических задач большой сложности. При выполнении теста МРТ с ограниченным выбором слушатель может использовать ограничения, присущие данной постановке задачи. Однако за пределами лаборатории слушателю редко задаются ограничения на задачу. Нет простого или прямого метода оценки по результатам теста МРТ с ограниченным выбором правильности распознавания речи оператором в ситуациях с меньшими ограничениями. Напротив, оценку систем с речевым выходом необходимо проводить при тех же требованиях к задаче, какие ожидаются в реальном практическом применении. Лабораторные эксперименты ведутся с целью сравнения различных систем. Они окажутся весьма полезными для разработки и применения систем, если дополнить их другими релевантными данными.

На основе результатов наших работ по восприятию синтезированной речи нам удалось сформулировать некоторые ограничения на использование систем с речевым выходом. Однако предстоит выполнить большой объем исследований. Нужны фундаментальные исследования, чтобы понять, какой эффект оказывают на восприятие синтезированной речи в сложной обстановке шум и искажения, как тренировка и ранее накопленный опыт влияют на восприятие, и как именно воспринимаемая естественность речи взаимодействует с ее разборчивостью. Сейчас, когда техника автоматического синтеза речи по тексту отработана до такой степени, что уже имеется несколько конкурентоспособных промышленных вариантов таких систем, в том числе и на русском языке, необходимы дальнейшие исследования, которые позволят нам понять как потенциальные возможности, так и недостатки систем с речевым выходом, а также, как осуществляется взаимодействие между человеком и этой новой техникой.

СПИСОК ЛИТЕРАТУРЫ

1. *Воронин В.М.* Восприятие и понимание естественной и синтезированной речи // Психологический вестник Уральского государственного университета. Выпуск 4. Екатеринбург: Издательство Уральского университета, 2003.
2. *Greene B.G., Manous L.M., Piosoni D.B.* Preliminary evolution of DECtalk // Speech Res. Lab. Tech. Note 84-03, Bloomington, IN, Indiana University, 1984.
3. *Nusbaum H.C., Piosoni D.B.* Some constraints on the perception of synthetic speech, Behavior Res. Methods instrum.1983.
4. *Nusbaum H.C., Piosoni D.B.* Perceptual evaluation of synthetic speech: Some constraints on the use of voice response systems // Proc. 3rd Voice Data Entry Systems Applications Conf. Sunnyvale, CA, Lockheed, 1983.
5. *Piosoni D.B.* Some measures of intelligibility and comprehension // Allen, S. Hunnicultt, and D.H. Klatt, Eds., Conversion of Unrestricted Text to Speech, Notes for MIT Summer Couse 6.69s, July 1979.
6. *Piosoni D.B.* Perception of speech: The human Listener as a cognitive interface // Speech Tecnol., pp. 10-23, 1982.
7. *Shiffrin R.M.* Capacity limitations in information processing attention, and memory // W.K. Estes, Ed., Handbook of Learning and Cognitive Processes, vol. 4, Hillsdale, Nj: Erlbaum, 1976.